SANDIA REPORT

SAND2007-6278
Unlimited Release
Printed September 2007

Interactomes to Biological Phase Space: a call to begin thinking at a new level in computational biology

W. Michael Brown and George S. Davidson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2007-6278
Unlimited Release
Printed September 2007

Interactomes to Biological Phase Space: a call to begin thinking at a new level in computational biology

W. Michael Brown and George S. Davidson
Computational Biology

Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1316

Abstract

Techniques for high throughput determinations of interactomes, together with high resolution protein collocalizations maps within organelles and through membranes will soon create a vast resource. With these data, biological descriptions, akin to the high dimensional phase spaces familiar to physicists, will become possible. These descriptions will capture sufficient information to make possible realistic, system-level models of cells. The descriptions and the computational models they enable will require powerful computing techniques. This report is offered as a call to the computational biology community to begin thinking at this scale and as a challenge to develop the required algorithms and codes to make use of the new data.

CONTENTS

FIGURES

Introduction

A detailed understanding of cells as systems is critical because the emergent phenomena of life is not disclosed by a simpler understanding of isolated biomolecules. Achieving these insights and mechanistic understandings will enable huge progress in the treatment of disease, alternative energy production, materials design, synthetic biochemistry, nanotechnology, and chem-bio warfare agent detection. However, we are still a long way from that goal.

We have only very incomplete answers to the problems that plague us. Why are there drastic differences between mice and humans, despite the overwhelming similarity in mouse and human genes? Why don't drugs that work on the same target in mice work in humans? Is drug toxicity due to non-specific drug interactions or naïve assumptions regarding the role of the target protein? Does the network of interacting molecules suggest a general target for cancer therapeutics?

These are not small holes in our knowledge; they reflect the need to undertake a vast program of research aimed at a fuller understanding of cells. We will measure our success in that endeavor by our ability to predict cellular responses across a broad range of conditions: for example, from normal to abnormal gene doses due to aneuploidy and chromosomal rearrangements and from normal physiological conditions to extreme environmental exposures and cytotoxic therapies. The required models and simulations are not yet achievable, in part because we lack the fundamental knowledge about which proteins collocate and interact with each other, but also because we lack the appropriate conceptual frameworks to deal with the complexity offered by whole cell mechanisms.

If we barely understand cells, how much further are we from understanding other areas of interest regarding the manipulation of life, from humans down to cells, and further down to protein catalysts? Human performance enhancement, battling the effects of aging, understanding cognition at a molecular level, metabolic engineering of cells for more efficient production of fuels, and protein engineering of novel catalysts are all limited by our current understanding of biology.

Unfortunately, merely extending the current level of protein annotations to the entire set of translated proteins is unlikely to break barriers on its own. We will still lack understanding of how these proteins interact to regulate biological processes and how the interplay of proteins and the morphology of organelles, cells, tissues, and organs ultimately enable life. Our ignorance can be lethal; engineering a protein to enhance performance within the cell for a particular reaction might ultimately be deleterious due to unknown regulatory functions provided by protein interactions. However, one can imagine a time when our understanding will greatly lower the risks. Stretching our imaginations even further, it is possible that whole biochemical industries that would otherwise rely on intact cells, might use smaller biomolecular machines if we can learn how to provide more stable and efficient cell-free processes. The excitement and power of bioengineering in its general sense seems palpable, but we must begin by taking small, achievable steps toward that new understanding of biology at a systems level.

Here, we present a review of the literature motivating the necessity of a systems approach. Although high-throughput genomics and proteomics approaches have made important progress, we show why further investigation into the role of proteins is required. We discuss important limitations based on the available data and computational methods. We discuss recent advances in experimental techniques that allow for an unprecedented amount of data to be collected regarding the complex roles of proteins in cellular processes. We discuss how the data can be used to:

1. Predict function for uncharacterized proteins
2. Identify new functions for known proteins
3. Group molecules into the organized machines responsible for cellular processes
4. Identify novel targets for disease and bioengineering
5. Understand molecular recognition and improve predictive methods
6. Facilitate protein engineering and cellular engineering.

We discuss the computational tools that will be required. Finally, we discuss the idea of a biological phase space and methods that can identify manifolds in this space in order to reduce the complexity of computational methods and enable new advances in bioengineering and the treatment of disease.

Obtaining a New Level of Understanding

The foundation of biochemistry and molecular biology was built by investigating protein functions in experiments involving only one or a few proteins at a time. This approach has identified seemingly straightforward pathways describing the influence of metabolic and signaling functions on biological processes. However, it has become increasingly clear that proteins function in complex networks and not in isolation.

Indeed, the investigation of isolated proteins has led to the identification of function for only a small fraction of the proteins predicted to exist in the human body. Additionally, it has become apparent that many proteins play multiple roles in distinct biological processes. In addition to the numerous signaling and nucleic acid binding proteins that have multiple functions and interact with multiple classes of partners, metabolic enzymes have also been shown to play additional roles [1, 2]. A significant number of proteins are found to be localized to multiple subcellular compartments; >28% are in two locations and >8% are in three locations according to one account [3] and this is likely an underestimate limited by known observations.

The complexity surrounding protein function, along with the idea of "molecular machines" responsible for biological processes, has led several researchers to describe genomics and proteomics as the discovery of a list of parts [4-6]. Beyond the parts, uncovering the schematic detailing how these proteins work together in biological processes is paramount for future investigations into the mechanism and treatment of disease, gene therapy, evolution, and metabolic engineering.

As a first step in this process, researchers have recently begun to tackle the notion of the *interactome* - the complete list of physical interactions mediated by all proteins of

an organism [7]. This understanding of proteins as part of a larger set of molecular machines and the communications between them promises a new biochemical framework for understanding processes as an integrated activity of highly interacting cellular components [8, 9]. Uncovering this interaction network has implications regarding the topology of biological interactions and will redefine how we look at protein function, evolution, and disease.

In discussing robustness and evolvability in living systems[10], Andreas Wagener has made the argument that evolution has left behind a tangle of prior, less efficient enzyme interactions in modern cells that act together with the more evolved enzymes. Understanding phenotypes and robustness to mutations and aberrant environmental conditions must take into account the remnants of these archaic interactions. He expects that it will be the case, unfortunately, that the increase in knowledge of the details on how cells work will coincide with increasingly opaque models – we will not be able to understandable what we know![1]

With that warning in mind, we recognize the need to collect the required data and deal with the complexity in new and novel ways, including one that we propose to call the *biological phase space*. We will discuss this concept later together with various mathematical techniques for dimensionality reduction that can describe the interplay of fundamental biomolecules within within that phase space. However, we will initially examine the problem with, perhaps, more familiar terminologies.

Network Topology

The topology of the regulatory network of biomolecules is relevant to investigations into network stability, dynamics and function, and approaches to reengineer biological processes [11, 12]. For example, the possibility that biological networks are scale-free could account for robustness in the face of mutation and environmental stress due to their resistance to random failure [12-16]. Proteins having many interactions in signaling pathways that have become deregulated in a cancer might represent more successful targets for therapeutics [12, 17, 18] when compared to heterogeneous targets identified from expression profiling. Additionally, there is debate as to whether highly-connected proteins are essential to survival [3, 14].

There is increasing evidence that a given biological function should not be assigned to a single protein, but rather that it emerges from the interaction of many components forming distinct functional modules [8, 9, 19-23]. For example, analysis of gene expression and protein interaction data for early embryogenesis in *Caenorhabditis elegans* revealed distinct groups of highly interconnected proteins with few or no links between them [24]. Analysis of interaction networks to isolate the densely connected subgraphs can identify distinct biological modules [25] and can suggest how the functions of these modules are regulated within the network.

[1] Personal communication between Andreas Wagner and George Davidson, in discussing his book, *Robustness and Evolvability* during the summer of 2006 at the University of New Mexico.

Evolution

Data on the biomolecular interaction network will allow for an understanding of evolution not only in terms of how a given mutation influences an isolated activity, but how the overall network is influenced [26-28]. For example, why are some cellular components conserved across species while others evolve rapidly [5, 29-31]? Analysis of protein interaction data has suggested that the evolutionary rate of a protein is not determined solely by its essentiality and individual fitness, but by its level of interaction with other proteins [32], the very topological motifs describing these interactions [31], and even preadaptation for activation by signals that will facilitate future interactions that have not yet evolved[33]. Additionally, analysis of several interacting protein pairs has revealed that they co-evolve [34, 35].

Genetic analysis has revealed that *Caenorhabditis elegans* and humans have a similar number of genes. This observation, together with the similarity between the sequences of human and mouse genes almost certainly suggests that species differences might not be due to individual functions of the component genes, but the complex interactions between them [36]. This broader interpretation of interacting genes may explain the surprising evidence that protein-protein interactions are not well-conserved across species [3, 37]; the story is in the complexity of many interactions, not in the specific pairwise interactions.

Uncovering the role of protein interaction networks in evolutionary processes will aid in our understanding of how perturbations due to disease, gene therapy, or metabolic engineering influence biomolecular processes as a whole. High throughput methods including expression arrays and whole genome sequences have demonstrated that preserved functional modules can be detected with clever algorithms and large databases[38-40]. Even greater insights should be expected when we are able to survey a large fraction of the proteome and are able to identify proteins with preserved collocation profiles across multiple tissues and species.

Very practically, the evolutionary requirement for preserving functional modules and their topologic features suggests an important criterion for the identification of drug targets in viral and bacterial pathogens prone to drug resistance. This approach might also be used to identify “linchpins” in protein networks describing oncogenic pathways [4].

Protein Function

At least 40% of human genes lack any functional annotation [41] and the amount of missing information due to the multiple roles of proteins is unknown. Large scale analysis of protein interaction networks is advantageous in that it allows for an unbiased inspection of protein function [25]. That is, investigation of a protein’s role within a specific biological context can give rise to misleading or incomplete information when viewed without regard to the “big picture”.

As an example where clues emerge from the big picture that otherwise be difficult to see with a tighter focus, consider the finding that proteins with similar functions are more interconnected by direct protein interactions than expected by chance [24] and that

proteins interacting with those involved in inherited disease are likely to cause similar disorders [3]. Interestingly, it has been shown that two proteins that have many interaction partners in common are more likely to be related biologically [42, 43]. Methods for identification of function for unknown proteins or protein complexes have been described based on their interactions in biomolecular networks [44-48] and such approaches have been utilized to identify functions for uncharacterized proteins involved in DNA repair and human cancer [49, 50].

Together with genomics and proteomics data, chromosome organization, protein-interaction networks and protein localization knowledge can be used to identify the function of unknown proteins, their role in molecular machines, and the regulation of these machines through network interactions. Additionally, identification of protein function within the biomolecular interaction network can shed new light on the mechanisms of drug toxicity; that is, toxicity might result not only from non-specific interactions with target proteins, but from our ignorance of how perturbations from drug action influence the network as a whole. In general, it is becoming clear that simple annotations of protein function, while intuitive to the biochemist and molecular biologist, are insufficient for describing the complicated roles of proteins in molecular networks. Advancing computational algorithms for querying the true effect of proteins on biological observables seems necessary for improving the success of future endeavors.

Disease

The value of a deep understanding of interactions, especially at a systems-level, is perhaps nowhere more obvious than with respect to diseases. In a recent study by Marc Vidal and colleagues, ~10% of the possible protein-interactions in the human genome were analyzed with yeast two-hybrid assays [25]. Indexing with the Online Mendelian Inheritance in Man database revealed 424 interacting pairs associated with human disease. Approximately, 77% of these interactions appeared to be new based on literature searches. With broader coverage of the human interactome, we can expect a great deal of medically valuable, new information. Additionally, by simultaneously investigating viral proteomes interacting with host proteomes during early infection, it is reasonable to anticipate the discovery of totally unexpected new targets for viral pathogens.

Interestingly, analysis of proteins mutated in inherited disorders has shown that they are likely to interact with proteins causing similar disorders, suggesting the existence of disease subnetworks[3]. Clarification of the nature of these disease networks may offer a novel method for identification of new disease targets. In addition to identifying new individual targets for disease, uncovering the protein interactions networks involved in the deregulation of diseases such as cancer can offer deeper biological insight into the development of therapeutics.

In an analogy given by Daniel Rhodes [4], traditional approaches for identifying new cancer targets are “akin to asking what makes an airplane different from an automobile, taking both apart, making a list of differences in the parts, and then focusing on a single part”. In contrast, systems-level approaches offer the potential for novel treatments to address the molecular heterogeneity and could provide needed insight into the difficulties in extrapolating from mouse models to human therapeutics. Here, the

systems approach idea is to simplify cancer signatures into coordinately regulated processes, transforming expression profiles into multidimensional interaction networks [4]. Such approaches can reveal common therapeutic targets for oncogenic pathways that might be missed due to the myriad of genes identified from expression profiling in a single pathway.

Limitations of Current Approaches

Aren't we already making progress? Is there really a need for a mid-course correction? Well, the answer is 'yes' to both questions. Large scale screens have been conducted, but they are not, yet, large enough. They are still so incomplete that conclusions drawn from the available samples are subject to distortions[11]. Further, cells are dynamical systems and the current high throughput methods hardly yield useful kinetics. A mid-course correction to our techniques and scientific goals could allow us to address the problems discussed below, and might yield the data required for useful cell models and simulations.

Various experimental approaches have been developed for identification of protein interactions including yeast two-hybrid assays, mass spectrometry of isolated protein complexes, protein chips, and hybrid approaches. Large-scale yeast two-hybrid screens have been performed for *Helicobacter pylori* [51], *Saccharomyces cerevisiae* [52, 53], *Drosophila melanogaster* [54]*, Caenorhabditis elegans* [43]*,* and humans [25, 55]. Interactome maps are currently available in databases containing interactions from literature curation, high-throughput experimental assays, and cross-species predictions based on "interologs". These databases include the Human Protein Reference Database [56], Biomolecular Interaction Network Database [57], Database of Interacting Proteins [58], Munich Information Center for Protein Sequences [59], Molecular Interactions Database [60], and IntAct [61].

Despite these efforts, current interactome maps have sampled only a small fraction of true interactions [6, 43, 54] and the overlap between interaction pairs obtained from distinct approaches is frighteningly low [41]. These deficiencies have important implications for the global analysis of biological networks due to the distorting impact that sampling can have on network topology [11]. For example, the scale-free topologies of partial networks should not imply a scale-free topology for the entire network [62]. In addition to missing data, the quality of existing data requires consideration. Reproducibility rates from yeast two-hybrid have been reported at ~55% for human interactome [25] and false positive rates have been reported to be ~50% [63]. False negative rates have been estimated as high as 85% for yeast two-hybrid and 50% for coaffinity purification with mass spectroscopy [63, 64]. Clearly, there is an outstanding need for alternative high-throughput approaches and additional data on protein binary interactions [7].

The limitations of yeast two-hybrid and affinity purification approaches have led some researchers to restrict the definition of the interactome to the "complete collection of binary protein-protein interactions detectable in one or more exogenous assay", without the consideration of the dynamic or functional properties of these interactions [25]. However, it is important to realize that the cell is not a well-mixed isotropic

solution. The effects of protein subcellular localization, cell trafficking and nuclear shuttling, posttranslational modification and splice variants, and the influence of molecular crowding on thermodynamics are likely to have a dramatic influence on biomolecular interaction networks. The futility of model parameterizations based on mass action kinetics, in light of current experimental approaches, suggests alternatives are required. Ignoring the context dependence of protein interactions can result in misleading information for interactions that do not occur *in vivo* or within a specific cell line [4]. Additionally, it cannot be determined from binary protein interaction data alone whether proteins interact with their partners simultaneously or at different times and locations. Biological processes may occur differently depending on the organism, the organ, the tissue, the cell, or the sub-cellular compartment.

Clearly, the development of computational algorithms to handle the large quantities of data involved in dynamic, biomolecular networks will play a central role in future advances. Approaches will be needed to assess the confidence in high-throughput data from multiple sources, utilize information on protein interactions for problems in molecular recognition, determine protein functions, identify drug targets, and visualize information. Perhaps most importantly, predictive methods will need to be developed to facilitate successful bioengineering and drug development.

A Path Forward

Ultimately, the dynamics of the human interactome network will need to be uncovered in order to elucidate where and when interactions take place and how they are regulated [21, 25]. The importance of network remodeling in space and time was recently shown in the transforming growth factor-β pathway [65], where spatiotemporal analysis revealed considerable partner switching in protein interactions and the loss of numerous interactions upon signaling in exchange for interactions in Smad complexes. New approaches are needed in order to understand how protein interactions are organized at the scale of the whole cell, how information, energy and molecules flow through biological networks, and how global and local properties of complex molecular networks influence biological properties and lead to human disease.

Building on the Interactome

We have illustrated in Figure 1 an approach for investigating the biomolecular networks responsible for life and examples of technologies that these investigations would enable or enhance. The interactome is the starting point. If two proteins are present together, will they interact? Expanding on this information can help us to address important problems regarding molecular recognition. The ability to inhibit protein interactions or, alternatively, to create new interactions relies on our ability *to predict* how proteins will interact with the use of bioinformatics and biophysics approaches; these approaches require a detailed understanding of existing interactions. Additionally, the information regarding all of the interactions of a protein is necessary to perform truly rational design of novel proteins.

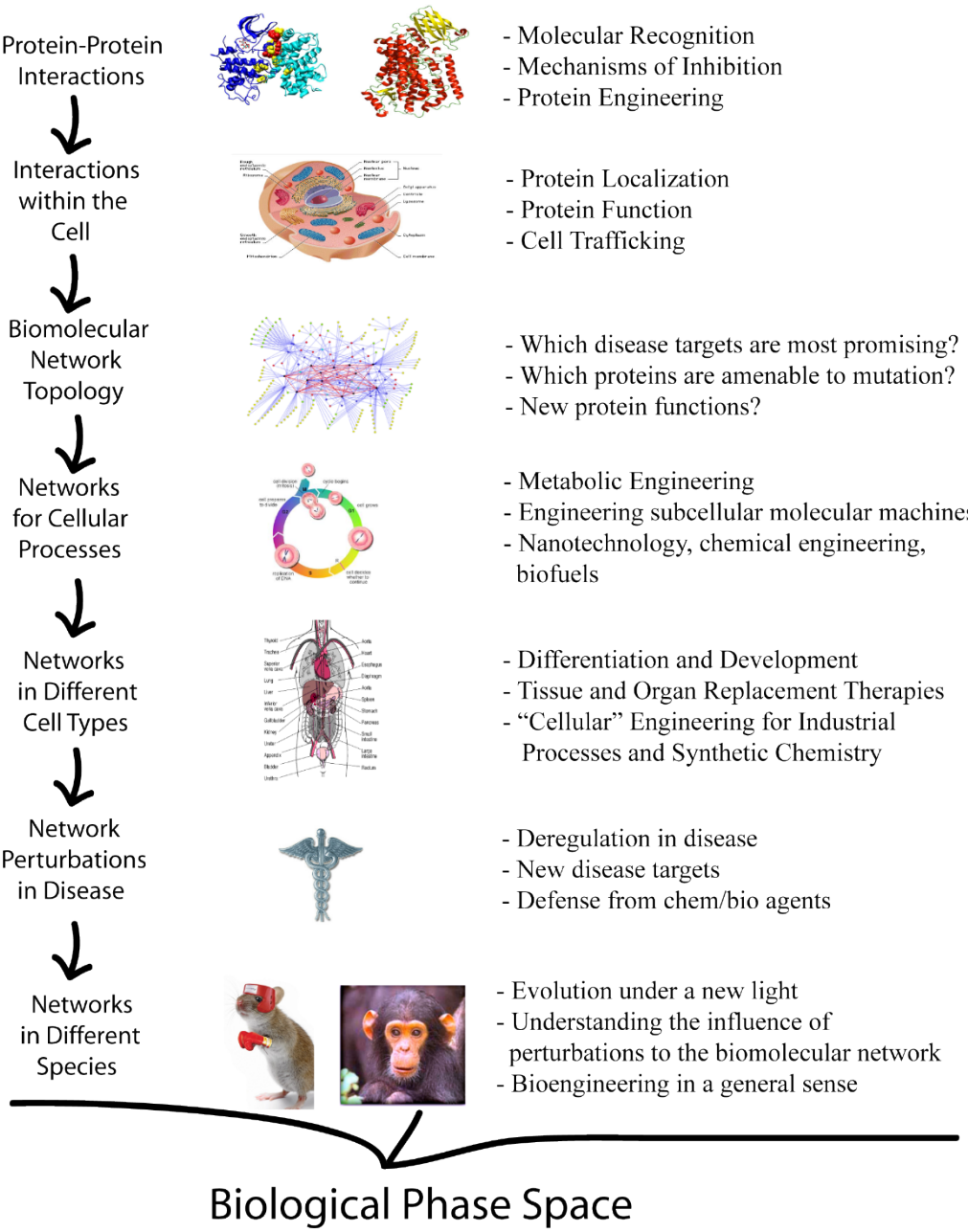

Figure 1. Illustration of the investigation into biomolecular networks at different levels and the technologies that would be enabled or enhanced. As a whole, this data facilitates a description of a biological phase space where the state of a biomolecular network maps to biological observables.

Automating High Throughput Collocalization Measurements

While information on binary *in vitro*-protein interactions is useful, understanding these interactions in the context of the cell cannot be achieved based only on this information. These interactions will depend on expression and localization within the cell and these two will depend on the type of cell and its internal information state. Therefore, a true understanding of biomolecular network topology will require the development of new high-throughput technologies for observing protein interaction networks as they occur in different states of a cell and within different types of cells. Obtaining this

information will give us insight into the mechanisms of life and disease along with a schematic for rational bioengineering.

Filling In the Interactome

Several recent breakthroughs are, indeed, enabling new approaches that can identify collocated proteins within their cellular compartments. A particularly interesting example is the work in Germany[66] where a laboratory with sixteen automated microscopes has been developed as a step along the pathway toward high throughput collocalization measurements. In their approach, robots implement serial fluorescent immuno-staining using flow cells on the microscope stages. They have demonstrated the ability to sequentially stain and image a large number of different epitopes within a tissue sample (or within individual cells under higher magnifications).

Figure 2 depicts how they measure the locations of several epitopes in a sequence and then combine these measurements to produce collocation maps. They reported results using about 100 different antibodies, and showed that various permutations of staining sequences gave similar results. Importantly, they are working to extend these methods to make measurements with a thousand different antibodies.

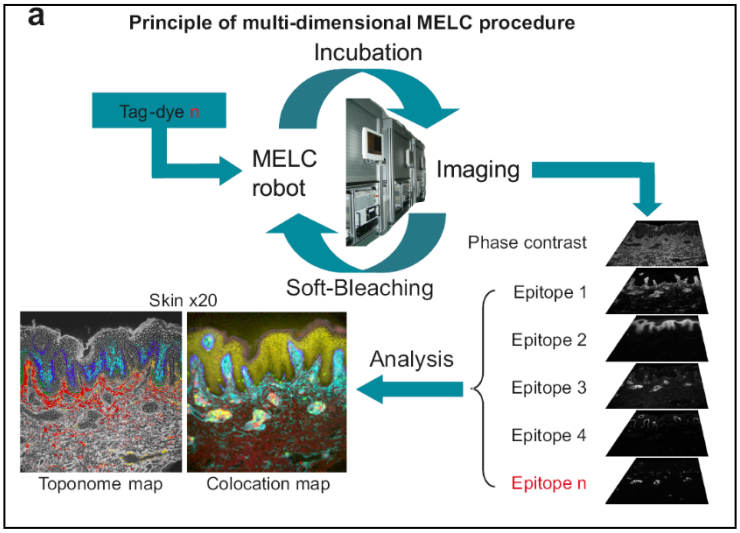

Figure 2 A depiction of the process implementing automated immuno-staining, followed by imaging and photo bleaching prior to the next round of staining and imaging. Note how images of individual epitopes can be combined with false coloring to render protein locations within the tissues, Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology [66] copyright 2006.

Optically based, high throughput proteomics systems with automated liquid sampling for the chemistries *can be seen as just a machine tool* for gathering the locations of hundreds if not thousands of proteins localized to their cellular organelles. Seen this way, it is clear how to scale the process in the same way that massively parallel super computers are made: combine and replicate a single common building block in a massively parallel way. For supercomputers the basic block is a simple, commercial computer, but for proteomics it would be something like that shown in. Figure 3. Scaling to a production facility with hundreds of these basic building blocks is very much like

any other manufacturing environment with many replicas of the basic machine tools. Like other manufacturing facilities, one must be concerned with the materials coming into the plant, the products being shipped, and the wastes being generated. With respect to that production stream, micromachined flow cells should give greater experimental control while minimizing the required chemicals and resulting waste.

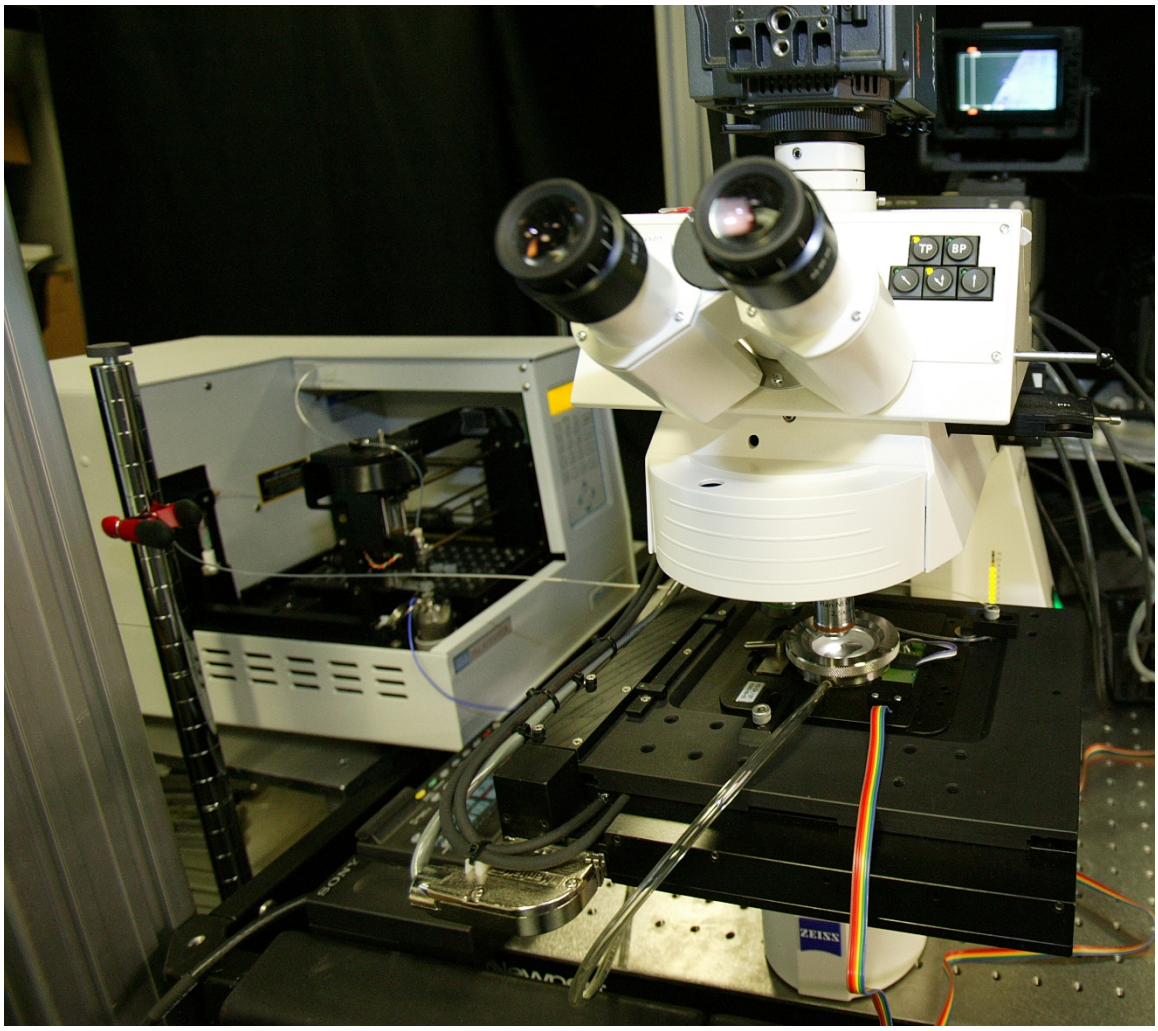

Figure 3 Automated microscopes with liquid sampling capability and flow cell stages capture the raw data, which must be further processed to identify collocations.

Micro-scale Reaction Vessels

The application of micromachined reaction vessels and microscopy is already yielding important new experimental methods, including single cell studies. In addition to methods for manipulating and imaging live cells, there are new opportunities for working with fixed cells. Consider, for example, a time course experiment after exposing groups to a treatment. The exposed cells could be sorted and moved into channels where they are rapidly fixed at the specified time points. Once fixed, the automated methods discussed above could be used to identify proteins collocating under the experimental conditions. By conducting as much of the experiment as possible within the micromachined

environment, human errors and problems with reproducibility should be greatly minimized.

Antibody Libraries

Of course one cannot use immuno-fluorescent staining without having the required antibodies. While many antibodies are available, the vast majority of proteins produced by cells from the species of most interest have not been isolated and used to produce antibodies. Further, the production of antibodies for proteins that are highly conserved between the target species and the animals used in antibody production remains problematic because they will not recognize the target species' proteins as different from their own, and no immune response will be invoked.

Certainly, phage display methods offer the promise of an alternative to traditional animal grown antibodies and could well play a critical role in high throughput optical proteomic approaches. In an alternate approach, David Peabody's work with MS2 bacteriophage[67] has led to another possible path toward the production of the required antibodies. MS2 phage can be induced to display antigenic peptides on their surface[68, 69]. These coat proteins can be grown with cell-free methods and will self assemble. Further, the dense, repetitive presentation of the polypeptides as part of the coat protein produces extremely strong immune responses, even breaking self immunity[70]. His methods should allow for the production of antibodies, even for those highly conserved proteins. It is expected that either phage-display or MS2 coat proteins can be developed to supply the required antibodies for very large fractions of the required proteomes and that these methods, together with the automated optical microscopes, can be used to construct interactomes of sufficient quality and coverage necessary for modeling.

Improving Optical Resolution of Protein Collocalizations

It should be noted that these high throughput optical proteomic methods remain subject to the diffraction limited resolution of traditional microscopy; light microscopy can only localize proteins to regions of about 200 nm. Finer resolution has traditionally required electron microscopy (EM) and immunostaining with attached gold beads, a relatively slow and expensive process subject to many experimental difficulties. However, very recent breakthroughs in single molecule imaging methods, combined with extensive computing, suggest a way to achieve resolution comparable to that of EM micrographs [71-73]. These methods are achieving resolutions of a few tens of nanometers, and should achieve even better resolutions in the future. Importantly, readily available fluorescent stains (Cy3/Cy5) have been shown to have the required optical switching properties[74, 75]; so, new exotic chemistries will not be required to collect the images. However, these images do require *a great deal of processing* to determine the protein locations, and further processing is needed to extract hypothetical networks from the collocation data.

Computing and New Machine Architectures

The combination of supercomputing with hundreds, or even thousands of automated microscopes using these single molecule imaging methods should enable the

identification of collocated proteins with resolutions sufficient to know if they are on the same side of an organelle membrane (resolutions better than 10 nm will be required). At these resolutions the confidence that two proteins are, or are not interacting should be greatly improved. However, the volume of data produced will be enormous, and a new generation of bioinformatics databases will be required. Fortunately, two new computer architectures suggest that it will be possible to deal efficiently with these data.

Massively parallel database machines like the Netezza Performance Server Model 10100 show that it will be possible to work with extremely large datasets (hundreds of terabytes, or more) at interactive rates. For instance, a recent test at the NIH/NCI Advanced Biomedical Computing Center showed speedups ranging from an order of magnitude up to twenty thousand times faster for a broad range of queries typical of their in-house work[2]. However, it is not sufficient to be able to look through large databases of reduced data; the ability to automatically find suggestive motifs will be critical. For example, identifying sets of a few molecules that generally seem to be collocated in and around a membrane until stimulated may be suggestive of a scaffolding for the initial stages of a signaling cascade. It will be important to be able to search for similar motifs under many different experimental conditions and in different species to find other such interactions or variants of the known interactions.

These motifs can be readily modeled as abstract graphs with nodes representing the proteins and arcs between nodes representing the interactions. Further, the nodes and arcs can have associated annotations describing the protein locations (which organelles, membrane bound, free in the cytosol, etc.) and the experimental conditions under which the interaction is observed (say, for example, during S-phase of the cell cycle after exposure to a particular toxin). Graph search algorithms are often extremely difficult, especially when the semantics of the graphs carry restrictions and temporal limitations as annotations. The resulting memory usage patterns greatly limit the value of caches and the use of multiple cores within high performance processors; essentially limiting the processor to an execution speed matching the slowest memory access rate. However, recent research at Sandia National Laboratories and elsewhere has had success using a new class of super computer that deals with the inherent distributed nature of graph data structures throughout the available memory[76].

Multithreaded dataflow machines are designed to deal with the inherent latency by maintaining many parallel threads of computation. Although every thread must wait for access to new data, on average one or more of the threads will have completed the required data access at any given point so that the processing unit never completely stalls. Tests on the Cray XMT and its predecessor have demonstrated large speedups over the performance possible with traditional supercomputers. These architectures and graph search algorithms are likely to play a large role in reducing the raw data from the high throughput optical proteomics facilities. The goal of such data fusion must be the automatic creation of interaction hypotheses and the inferences required to estimate network dynamics.

Together, automated microscopy, improved microscale reaction vessels, single molecule imaging methods, and new computing technologies are expected to be

[2] Personal communication from Todd Scofield, June 2007.

transformative for biology. A production capability should yield a new kind of data that is unprecedented in scale, depth, and coverage of biological conditions. These data will enable whole cell and tissue level understanding of systems biology that will elucidate information and energy flow through biological networks and may enable computational modeling of the global and local properties of complex molecular networks. This modeling at high levels of detail can, in turn, move us toward predictive cell science and enable the engineering of desired biological properties and new therapies for human disease.

Having suggested that we are on the threshold of having all of the elements required to measure and analyze *in vivo* protein interactions, and having seen that other research threads are developing the new tools required to search through and analyze the raw data from those methods, we must now consider the tools required for data analysis, modeling, and simulation.

Data Fusion and Uncertainty Quantification

There is concern with any high-throughput experimental approach about the quality of the data. With yeast two-hybrid approaches, this concern is at the forefront due to high rates of false positives and false negatives. Therefore, methods for validating data, assessing self-consistency, and metrics for confidence must be applied in concert with high-throughput approaches. For binary "interactome-style" data, numerous methods have already appeared in the literature. For example, protein interactions detected in multiple assays are more likely to be true positives, especially if the assays utilize distinct methodological techniques [7, 63, 77, 78]. Methods integrating data from genomics/transcriptomics (coexpression), proteomics (collocalization), and interactomics (binary interaction data) can be integrated to provide high confidence data [49, 79-86]. Confidence scoring systems have been developed that calculate the likelihood of an interaction using various parameters including attributes of the proteins and the specific assays, whether the interaction was detected by other technologies or screens, and network topology [42, 54, 87-90]. For example, genes with similar expression profiles are more likely to encode interacting proteins [24]. In general, statistical models must be integrated with existing knowledge in order to improve confidence predictions. For example, there is evidence for early expression and later use widely throughout the cell cycle in yeast [91] and evidence for sequestration of mRNAs for rapid release under signal triggered conditions [92], both of which could confound statistical analyses.

These approaches for data-fusion extend beyond applications for assessing confidence and validating data. For example, many observations have been made that allow for an integrated approach for determining the function of uncharacterized proteins. Genes with similar expression profiles are more likely to show enriched phenotype correlation and genes with phenotypic similarity are more likely to encode proteins that interact with each other [24, 91]. Additionally, bacterial genomes are known to be organized into regions that tend to code for proteins with similar functions and correlation is enriched with organizations that are conserved across different species [93]. The adjacency of genes in various bacterial genomes has been used to predict functional relationships between the corresponding proteins [94, 95].

These observations, together with spatiotemporal analysis of protein interactions can provide high confidence assessments of the complicated roles of proteins in molecular biology and biochemistry. Ultimately, methods for data fusion and uncertainty quantification (see for example,[96]) should play an important role in unraveling the intricacies of biomolecular interaction networks and provide a systems-level integration with the large amounts of data generated from genomics, proteomics, and metabolomics approaches.

Protein Interaction Domains and Molecular Recognition

Understanding molecular recognition, how biomolecules interact, is critical for methods that attempt to inhibit interactions or create new ones, for example in a new cancer therapy. While both physics-based [97] and informatics-based [98-102] approaches have been developed for identifying protein interactions, both approaches require data for empirical training and, currently, each suffers from low general accuracies. In the former, information on protein interactions is required for parameterizing force-fields capable of distinguishing low energy protein-protein configurations from the high energy ones that are unlikely to result in stable interactions. In the latter, information on observed protein interactions is used to train models that predict interactions and the interaction domains on the proteins involved.

Expanding the database of protein-protein interactions with increased sampling in different cell types, cell processes, tissues, and species is necessary for improving the accuracy of methods that attempt to predict or modulate protein interactions. This information in turn is necessary for problems in bioengineering and medicine. As we accumulate and understand these data, we will have the ability to engineer proteins with enhanced catalytic functions while maintaining their roles in biomolecular networks. Perhaps, we will even be able to engineer proteins leading to new biomolecular networks with desired properties or synthesize inhibitors to protein interactions involved in oncogenic and disease pathways.

Network Topology, Graph Theory, Clustering, and Visualization

Graph theoretic approaches are useful for interpreting the large amount of data generated by high-throughput approaches. They can identify the topology of biomolecular networks, sub-graphs representing functionally organized groups of proteins and their roles as molecular machines, drug targets, and targets for bioengineering. Methods have already been developed for identifying densely connected subgraphs [103], functionally enriched complexes [104], for unsupervised clustering and visualization [105], and for dividing biomolecular networks into modules [21, 87, 106-109]. While these methods create a foundation for analyzing biomolecular networks, they have all focused on binary interaction data. Many authors have used these approaches for identifying global properties of biomolecular networks, however, these are of nebulous value because they neglect interactions in a context relevant to the cell. Future methods will need to incorporate data on when, where, and why interactions occur in order to be of value.

In addition to data analysis, methods will be needed to provide an interface to the experimentalist. One might imagine an interface where the user enters a protein identifier and obtains a coordinated story about the protein and its cellular roles. This story might be synthesized on-the-fly with text, figures, and even video describing dynamic changes. Such a story would include, information on the molecules interacting with the specified protein, metabolic functions, locations within the cell, roles in various cellular processes and tissues, domains responsible for various function, evolutionary changes that have occurred and how these have affected the biological system as whole, and pictures of the protein in overall biomolecular networks.

Recasting the Problem and Biological Phase Space

Obtaining information on known biomolecular networks is one thing; using this information for bioengineering and medicine is another. Although the value of knowledge on biomolecular networks is not debated, as far as its role in providing starting points for experimental endeavors, predictions of the true effects of perturbations to the network are required for rational design. It is this ability that will facilitate the transition from knowledge to unhindered technological breakthroughs in the many scientific disciplines spanned by the biosciences and bioengineering. The complexity involved in predictive methods is daunting, however. Certainly, an *ab initio* approach utilizing quantum mechanics to simulate the biomolecular network will remain beyond our reach. We will need, instead, simpler models that can be realized with the available computing resources. However, higher level formulations utilizing mass-balance kinetics do tradeoff computational complexity against the requirement for extensive experimental data. It therefore seems intractable to describe biological observables with a phase space parameterized by molecular concentrations and stoichiometric rules for molecular interactions and chemical transitions.

One alternative is to recast the problem and to adopt a new definition of biological phase space in terms of data *that can be obtained*. For example, consider a phase space where biological observables are described in terms of protein sequence data from genomics, expression levels from transcriptomics, and data on biomolecular interaction networks as obtained in different cellular processes, tissues, and species. Although the dimensionality of the space is still very high, there is a kind of order within which lies the very motivation for this work – biomolecules are all connected in an intricate network. *Life does not exist as a uniform sampling in this phase space, but as a manifold with lower dimensionality.*

We can relate this description back to the failure of rational drug design that was presented earlier in this report. In that example, we posited a protein and its function, which was assumed to be involved in a disease. That function might only be a single dimension in the biological phase space. When we limit ourselves to thinking only along this single dimension, we are surprised to find that the drug and its single inhibitory function ultimately fails to become a useful therapy. The odds of developing a successful drug from identification of targets to clinical trials is incredibly small and has left many serious diseases such as diabetes and leukemia with few treatment options despite the huge resources given to scientists working in these fields. We suggest that identifying

more inclusive manifolds can shed new light on our endeavors, and will lead to more successes in health care and the engineering of biologically important molecules.

Recently, novel approaches for identifying non-linear manifolds in high dimensional spaces have appeared in the literature [110-112]. These approaches have been applied to problems such as protein folding in traditional phase spaces and are applicable to problems in the biological domain. Dimensionality reduction can be utilized to tie together samples from different cell types and different species into a manifold giving equations that relate intricate networks of biomolecules to biological observables.

Returning to the drug design just discussed, these methods may allow a broader, but still practical design methodology. Rather than blindly attacking a protein in drug design or bioengineering, we can obtain information on how inhibiting a protein will affect the system as a whole based on the natural changes that have occurred through evolution and are present in different cell types. With this intervention, we might invoke a series of treatments as preparation for the definitive drug. The initial manipulations would be used to move the cells and tissues toward regions in their phase space where they would be more likely to be successfully treated. While analysis of the large amounts of data necessary to develop this approach does present challenges, advances in high performance computing, along with the ability to parallelize current algorithms in terms of dense matrix decompositions and linear algebra solvers, can provide a solution. In addition to predicting the effects of perturbations to biomolecular networks, these approaches offer the potential for uncovering high-level rules that can facilitate spatiotemporal modeling of the dynamics of biomolecules composing entire cells on modern supercomputers.

Summary

It has become increasingly apparent that biomolecules act in a complicated temporal and spatial network rather than in isolation. Understanding these inter-related components is critical for future endeavors in bioengineering and medicine. Researchers have made the first steps in this direction with the development of high-throughput methods for identifying protein-protein interactions and pictures of static interaction networks. In theory, mass balance kinetic models might utilize this information for simulating how network perturbations influence biological observables. In practice, that approach is unattainable due to the overwhelming quantity of experimental data required, disparity in experimental conditions used to make measurements, the unknown effects of missing and uncertain data on experimental results, and an incomplete understanding of the effects of cell trafficking and morphology on kinetic models. This limitation precludes an accurate analysis of how these protein interactions regulate cellular processes; which is, of course, necessary if we are to understand how these networks can be modified to achieve new objectives. Therefore, future efforts must focus on protein interactions in a context relevant to cellular processes with high-throughput identification of the dynamics of protein interactions within the cell. This approach, abstracted from traditional chemical dynamics, could lead to predictive models and to accurate simulations that will augment our intuition and understanding of how the biomolecular networks regulate cellular processes and how they, ultimately, perform tissue specific functions.

Methods for assembling the data required for these new approaches are rapidly developing. It is important that the bioinformatics and cell modeling communities not wait, but begin immediately to plan how they can best use these new data.

References

1. Jeffery, C.J., *Moonlighting proteins: old proteins learning new tricks.* Trends in Genetics, 2003. **19**(8): p. 415-417.
2. Sriram, G., et al., *Single-gene disorders: What role could moonlighting enzymes play?* American Journal of Human Genetics, 2005. **76**(6): p. 911-924.
3. Gandhi, T.K.B., et al., *Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.* Nature Genetics, 2006. **38**(3): p. 285-293.
4. Rhodes, D.R. and A.M. Chinnaiyan, *Integrative analysis of the cancer transcriptome.* Nature Genetics, 2005. **37**: p. S31-S37.
5. Kitano, H., *Systems biology: A brief overview.* Science, 2002. **295**(5560): p. 1662-1664.
6. Uetz, P. and R.L. Finley, *From protein networks to biological systems.* Febs Letters, 2005. **579**(8): p. 1821-1827.
7. Vidal, M., *Interactome modeling.* Febs Letters, 2005. **579**(8): p. 1834-1838.
8. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761): p. C47-C52.
9. Oltvai, Z.N. and A.L. Barabasi, *Life's complexity pyramid.* Science, 2002. **298**(5594): p. 763-764.
10. Wagner, A., *Robustness and Evolvability in Living Systems*. Princeton Studies in Complexity. 2005, Princeton, NJ: Princeton University Press. 383.
11. Han, J.D.J., et al., *Effect of sampling on topology predictions of protein-protein interaction networks.* Nature Biotechnology, 2005. **23**(7): p. 839-844.
12. Barabasi, A.L. and Z.N. Oltvai, *Network biology: Understanding the cell's functional organization.* Nature Reviews Genetics, 2004. **5**(2): p. 101-U15.
13. Albert, R., H. Jeong, and A.L. Barabasi, *Error and attack tolerance of complex networks.* Nature, 2000. **406**(6794): p. 378-382.
14. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-42.
15. Sun, S., et al., *Error and attack tolerance of evolving networks with local preferential attachment.* Physica a-Statistical Mechanics and Its Applications, 2007. **373**: p. 851-860.
16. Wagner, A., *Robustness against mutations in genetic networks of yeast.* Nature Genetics, 2000. **24**(4): p. 355-361.
17. Apic, G., et al., *Illuminating drug discovery with biological pathways.* Febs Letters, 2005. **579**(8): p. 1872-1877.
18. Vogelstein, B., D. Lane, and A.J. Levine, *Surfing the p53 network.* Nature, 2000. **408**(6810): p. 307-310.
19. Snel, B., P. Bork, and M.A. Huynen, *The identification of functional modules from the genomic association of genes.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(9): p. 5890-5895.
20. Bork, P., et al., *Protein interaction networks from yeast to human.* Current Opinion in Structural Biology, 2004. **14**(3): p. 292-299.
21. Han, J.D.J., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature, 2004. **430**(6995): p. 88-93.
22. Alberts, B., *The cell as a collection of protein machines: Preparing the next generation of molecular biologists.* Cell, 1998. **92**(3): p. 291-294.

23. Spirin, V. and L.A. Mirny, *Protein complexes and functional modules in molecular networks.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(21): p. 12123-12128.
24. Gunsalus, K.C., et al., *Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis.* Nature, 2005. **436**(7052): p. 861-865.
25. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-1178.
26. Eisenberg, E. and E.Y. Levanon, *Preferential attachment in the protein network evolution.* Physical Review Letters, 2003. **91**(13).
27. Pereira-Leal, J.B., et al., *An exponential core in the heart of the yeast protein interaction network.* Molecular Biology and Evolution, 2005. **22**(3): p. 421-425.
28. Qin, H., et al., *Evolution of the yeast protein interaction network.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(22): p. 12820-12824.
29. Hasty, J., D. McMillen, and J.J. Collins, *Engineered gene circuits.* Nature, 2002. **420**(6912): p. 224-230.
30. Rao, C.V., D.M. Wolf, and A.P. Arkin, *Control, exploitation and tolerance of intracellular noise.* Nature, 2002. **420**(6912): p. 231-237.
31. Wuchty, S., Z.N. Oltvai, and A.L. Barabasi, *Evolutionary conservation of motif constituents in the yeast protein interaction network.* Nature Genetics, 2003. **35**(2): p. 176-179.
32. Fraser, H.B., et al., *Evolutionary rate in the protein interaction network.* Science, 2002. **296**(5568): p. 750-752.
33. Bridgham, J.T., Sean M. Carroll, and J.W. Thornton, *Evolution of Hormone-Receptor Complexity by Molecular Exploitation.* Science, 2006. **312**(97): p. 97-101.
34. Fryxell, K.J., *The coevolution of gene family trees.* Trends in Genetics, 1996. **12**(9): p. 364-369.
35. Pages, S., et al., *Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: Prediction of specificity determinants of the dockerin domain.* Proteins-Structure Function and Genetics, 1997. **29**(4): p. 517-527.
36. Valencia, A. and F. Pazos, *Computational methods for the prediction of protein interactions.* Current Opinion in Structural Biology, 2002. **12**(3): p. 368-373.
37. Mika, S. and B. Rost, *Protein-protein interactions more conserved within species than across species.* Plos Computational Biology, 2006. **2**(7): p. 698-709.
38. Kim, S.K., et al., *A gene expression map for Caenorhabditis elegans.* Science, 2001. **293**(5537): p. 2087-2092.
39. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-255.
40. Srinivasan, B.S., et al., *Functional genome annotation through phylogenomic mapping.* Nature Biotechnology, 2005. **23**(6): p. 691-698.
41. Cusick, M.E., et al., *Interactome: gateway into systems biology.* Human Molecular Genetics, 2005. **14**: p. R171-R181.
42. Goldberg, D.S. and F.P. Roth, *Assessing experimentally derived interactions in a small world.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(8): p. 4372-4376.

43. Li, S.M., et al., *A map of the interactome network of the metazoan C-elegans.* Science, 2004. **303**(5657): p. 540-543.
44. Bowers, P.M., et al., *Use of logic relationships to decipher protein network organization.* Science, 2004. **306**(5705): p. 2246-2249.
45. Hishigaki, H., et al., *Assessment of prediction accuracy of protein function from protein-protein interaction data.* Yeast, 2001. **18**(6): p. 523-531.
46. Oliver, S., *Guilt-by-association goes global.* Nature, 2000. **403**(6770): p. 601-603.
47. Schwikowski, B., P. Uetz, and S. Fields, *A network of protein-protein interactions in yeast.* Nature Biotechnology, 2000. **18**: p. 1257-1261.
48. Vazquez, A., et al., *Global protein function prediction from protein-protein interaction networks.* Nature Biotechnology, 2003. **21**(6): p. 697-700.
49. Boulton, S.J., et al., *Combined functional genomic maps of the C-elegans DNA damage response.* Science, 2002. **295**(5552): p. 127-131.
50. Rhodes, D.R., et al., *Probabilistic model of the human protein-protein interaction network.* Nature Biotechnology, 2005. **23**(8): p. 951-959.
51. Rain, J.C., et al., *The protein-protein interaction map of Helicobacter pylori.* Nature, 2001. **409**(6817): p. 211-215.
52. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(8): p. 4569-4574.
53. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-627.
54. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-1736.
55. Colland, F., et al., *Functional proteomics mapping of a human signaling pathway.* Genome Research, 2004. **14**(7): p. 1324-1332.
56. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Research, 2003. **13**(10): p. 2363-2371.
57. Bader, G.D., D. Betel, and C.W.V. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Research, 2003. **31**(1): p. 248-250.
58. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Research, 2004. **32**: p. D449-D451.
59. Pagel, P., et al., *The MIPS mammalian protein-protein interaction database.* Bioinformatics, 2005. **21**(6): p. 832-834.
60. Zanzoni, A., et al., *MINT: a Molecular INTeraction database.* Febs Letters, 2002. **513**(1): p. 135-140.
61. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Research, 2004. **32**: p. D452-D455.
62. Stumpf, M.P.H., C. Wiuf, and R.M. May, *Subnets of scale-free networks are not scale-free: Sampling properties of networks.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(12): p. 4221-4224.
63. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
64. Edwards, A.M., et al., *Bridging structural biology and genomics: assessing protein interaction data with known complexes.* Trends in Genetics, 2002. **18**(10): p. 529-536.

65. Barrios-Rodiles, M., et al., *High-throughput mapping of a dynamic signaling network in mammalian cells.* Science, 2005. **307**(5715): p. 1621-1625.
66. Schubert, W., et al., *Analyzing proteome topology and function by automated multidimensional fluorescence microscopy.* Nature Biotechnology, 2006. **24**(10): p. 1270-1278.
67. Peabody, D.S., *Subunit fusion confers tolerance to peptide insertions in a virus coat protein.* Archives of Biochemistry and Biophysics, 1997. **347**(1): p. 85-92.
68. van Meerten, D., et al., *Peptide display on live MS2 phage: restrictions at the RNA genome level.* Journal of General Virology, 2001. **82**: p. 1797-1805.
69. Brown, W.L., et al., *RNA bacteriophage capsid-mediated drug delivery and epitope presentation.* Intervirology, 2002. **45**(4-6): p. 371-380.
70. Chackerian, B., *Virus-like particles-flexible platforms for vaccine development.* Expert Review of Vaccines, 2007. **6**(3): p. 381-390.
71. Betzig, E., et al., *Imaging intracellular fluorescent proteins at nanometer resolution.* Science, 2006. **313**(5793): p. 1642-1645.
72. Rust, M.J., M. Bates, and X.W. Zhuang, *Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM).* Nature Methods, 2006. **3**(10): p. 793-795.
73. Hell, S.W., *Far-field optical nanoscopy.* Science, 2007. **316**(5828): p. 1153-1158.
74. Bates, M., T.R. Blosser, and X.W. Zhuang, *Short-range spectroscopic ruler based on a single-molecule optical switch.* Physical Review Letters, 2005. **94**(10): p. 108101.
75. Heilemann, M., et al., *Carbocyanine dyes as efficient reversible single-molecule optical switch.* Journal of the American Chemical Society, 2005. **127**(11): p. 3801-3806.
76. Berry, J., et al., *Software and algorithms for graph queries on multithreaded architectures*, in *Workshop on Multithreaded Architectures and Applications 2007*. 2007: Long Beach, CA.
77. Ge, H., A.J.M. Walhout, and M. Vidal, *Integrating 'omic' information: a bridge between genomics and systems biology.* Trends in Genetics, 2003. **19**(10): p. 551-560.
78. Walhout, M., et al., *A model of elegance.* American Journal of Human Genetics, 1998. **63**(4): p. 955-961.
79. Begley, T.J., et al., *Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping.* Molecular Cancer Research, 2002. **1**(2): p. 103-112.
80. Ge, H., et al., *Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.* Nature Genetics, 2001. **29**(4): p. 482-486.
81. Grigoriev, A., *A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae.* Nucleic Acids Research, 2001. **29**(17): p. 3513-3519.
82. Jansen, R., D. Greenbaum, and M. Gerstein, *Relating whole-genome expression data with protein-protein interactions.* Genome Research, 2002. **12**(1): p. 37-46.
83. Kemmeren, P., et al., *Protein interaction verification and functional annotation by integrated analysis of genome-scale data.* Molecular Cell, 2002. **9**(5): p. 1133-1143.

84. Walhout, A.J.M., et al., *Integrating interactome, phenome, and transcriptome mapping data for the C-elegans germline.* Current Biology, 2002. **12**(22): p. 1952-1958.
85. Lee, I., et al., *A probabilistic functional network of yeast genes.* Science, 2004. **306**(5701): p. 1555-1558.
86. Zhang, L.V., et al., *Predicting co-complexed protein pairs using genomic and proteomic data integration.* Bmc Bioinformatics, 2004. **5**.
87. Bader, J.S., et al., *Gaining confidence in high-throughput protein interaction networks.* Nature Biotechnology, 2004. **22**(1): p. 78-85.
88. Deane, C.M., et al., *Protein interactions - Two methods for assessment of the reliability of high throughput observations.* Molecular & Cellular Proteomics, 2002. **1**(5): p. 349-356.
89. Saito, R., H. Suzuki, and Y. Hayashizaki, *Construction of reliable protein-protein interaction networks with a new interaction generality measure.* Bioinformatics, 2003. **19**(6): p. 756-763.
90. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic Acids Research, 2005. **33**: p. D433-D437.
91. Werner-Washburne, M., et al., *Comparative analysis of multiple genome-scale data sets.* Genome Research, 2002. **12**(10): p. 1564-1573.
92. Aragon, A.D., et al., *Functional differentiation of quiescent and non-quiescent cells in yeast stationary-phase cultures.* in revision (minor) for Molecular Biology of the Cell, 2007.
93. Tamames, J., et al., *Conserved clusters of functionally related genes in two bacterial genomes.* Journal of Molecular Evolution, 1997. **44**(1): p. 66-73.
94. Dandekar, T., et al., *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends in Biochemical Sciences, 1998. **23**(9): p. 324-328.
95. Overbeek, R., et al., *Use of contiguity on the chromosome to predict functional coupling.* In Silico Biology, 1999. **1**: p. 93-108.
96. Ayub, B. and G.J. Klir, *Uncertainty Modeling and Analysis in Engineering and the Sciences*. 1 ed. 2006, Boca Raton, FL: Chapman & Hall/CRC.
97. Smith, G.R. and M.J. Sternberg, *Prediction of protein-protein interactions by docking methods.* Curr Opin Struct Biol, 2002. **12**(1): p. 28-35.
98. Ben-Hur, A. and W.S. Noble, *Kernel methods for predicting protein-protein interactions.* Bioinformatics, 2005. **21 Suppl 1**: p. i38-46.
99. Deng, M., et al., *Inferring domain-domain interactions from protein-protein interactions.* Genome Res, 2002. **12**(10): p. 1540-8.
100. Jansen, R., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data.* Science, 2003. **302**(5644): p. 449-53.
101. Martin, S., D. Roe, and J.L. Faulon, *Predicting protein-protein interactions using signature products.* Bioinformatics, 2005. **21**(2): p. 218-26.
102. Sprinzak, E. and H. Margalit, *Correlated sequence-signatures as markers of protein-protein interaction.* J Mol Biol, 2001. **311**(4): p. 681-92.
103. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* Bmc Bioinformatics, 2003. **4**.
104. Berriz, G.F., et al., *Characterizing gene sets with FuncAssociate.* Bioinformatics, 2003. **19**(18): p. 2502-2504.

105. Sultan, M., et al., *Binary tree-structed vector quantization approach to clustering and visualizing microarray data.* Bioinformatics, 2002. **18**: p. S111-S119.
106. Dunn, R., F. Dudbridge, and C.M. Sanderson, *The use of edge-betweenness clustering to investigate biological function in protein interaction networks.* Bmc Bioinformatics, 2005. **6**.
107. Pereira-Leal, J.B., A.J. Enright, and C.A. Ouzounis, *Detection of functional modules from protein interaction networks.* Proteins-Structure Function and Genetics, 2004. **54**(1): p. 49-57.
108. Rives, A.W. and T. Galitski, *Modular organization of cellular networks.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(3): p. 1128-1133.
109. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nature Genetics, 2002. **31**(1): p. 64-68.
110. Das, P., et al., *Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(26): p. 9885-9890.
111. Hinton, G.E. and R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks.* Science, 2006. **313**(5786): p. 504-507.
112. Donoho, D.L. and C. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(10): p. 5591-5596.

Distribution:

1	MS-0123	LDRD, 1011 (electronic copy)
1	MS-1316	W. Michael Brown, 1412 (electronic copy)
1	MS-1316	George S. Davidson, 1412 (electronic copy)
1	MS-0899	Technical Library, 9536 (electronic copy)